GPO PRICE $ _____

CFSTI PRICE(S) $ _____

Hard copy (HC) 2.00

Microfiche (MF) .50

ff 653 July 65

UNLIMITED SIMULTANEOUS DISCRIMINATION INTERVALS IN REGRESSION

by

G. J. Lieberman, R. G. Miller, Jr., and M. A. Hamilton

TECHNICAL REPORT NO. 90

July 29, 1966

Supported by the Army, Navy, Air Force, and NASA under
Contract Nonr-225(53) (NR-042-002)
with the Office of Naval Research

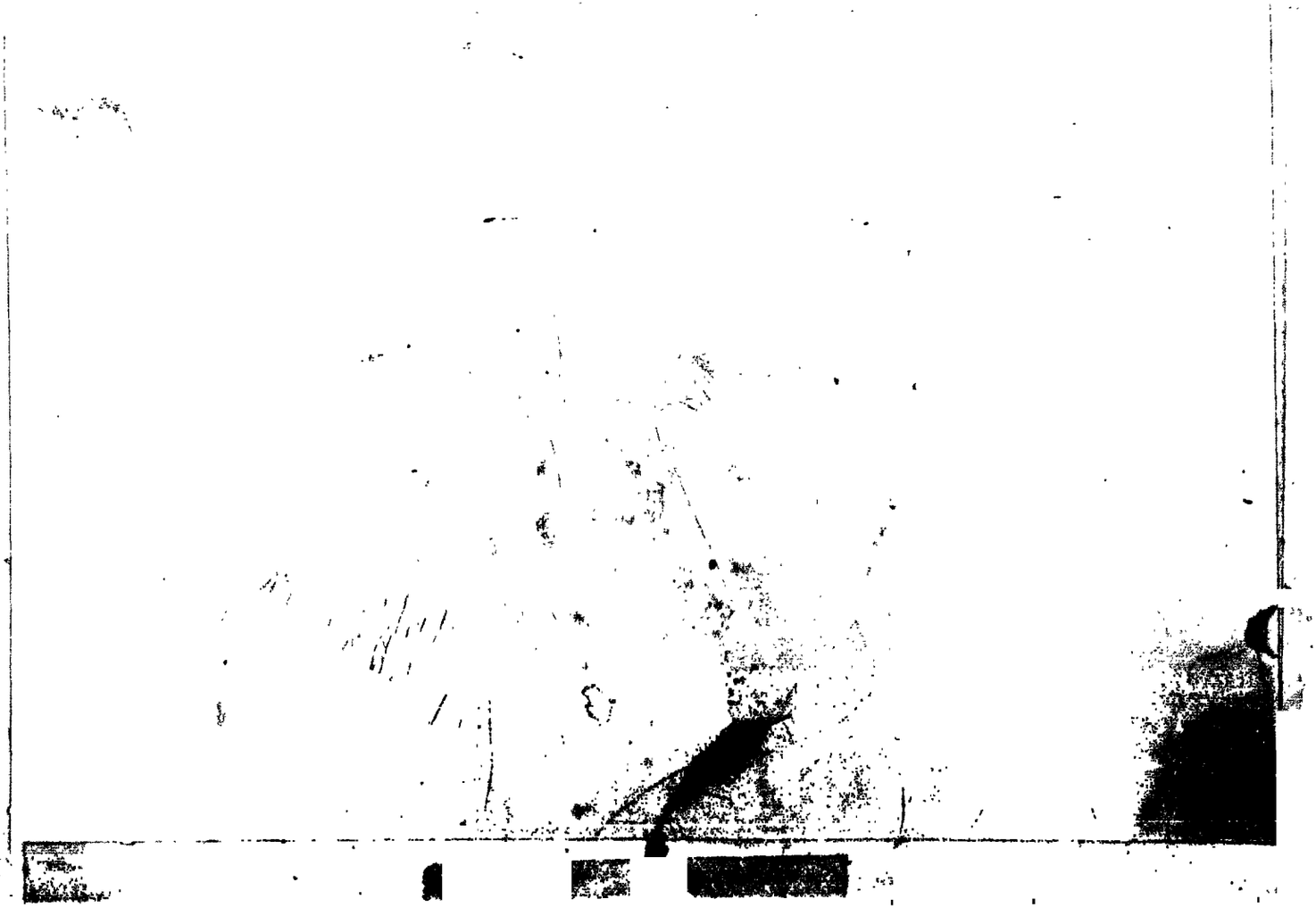
Gerald J. Lieberman, Project Director

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

BLANK PAGE



UNLIMITED SIMULTANEOUS DISCRIMINATION INTERVALS IN REGRESSION

by

G. J. Lieberman, R. G. Miller, Jr., and M. A. Hamilton

I. Introduction

In a recent paper [2] Lieberman and Miller considered the problem of obtaining simultaneous tolerance intervals in regression. Such intervals are easily described for the simple linear regression model $E(Y) = \alpha + \beta x$. Tolerance intervals $[L_x(P), U_x(P)]$ were found which are based upon the same estimated linear regression and which have the property that with confidence $1-\alpha$ the interval $[L_x(P), U_x(P)]$ contains $100P$ percent of the normal distribution centered at $\alpha + \beta x$ for any x and any P. In terms of predictions, a future value of Y at level x of the independent variable will lie in $[L_x(P), U_x(P)]$ with probability at least P with confidence $1-\alpha$ for any x and any P. The proper interpretation of this probability statement is that the "confidence $1-\alpha$" refers to the sample from which the regression line is estimated and the "probability at least P" to the sampling distribution of future observations. If for a single regression line one asserts that the proportion of future observations falling within the given tolerance limits (for any x) is at least P, and similar statements are made repeatedly for different regression lines, then for $100(1-\alpha)$ percent of the different regression lines the statements will be correct.

Simultaneous tolerance intervals have a different interpretation than that given for prediction intervals. Lieberman[1] treated the problem of determining the joint prediction intervals for the future

values of Y at each of K(known) separate settings of the independent variable x, when all K predictions are based upon the original fitted regression line. A probability statement concerning this joint prediction interval containing the K future values of Y was given.

An experimenter is sometimes interested in joint prediction intervals, and sometimes interested in simultaneous tolerance intervals. When the number K or prediction intervals is large, the resultant intervals may be useless. In other cases, the number K may be unknown so that the prediction intervals may not be applicable. For these cases, as well as those situations when the experimenter actually wants a tolerance interval instead of a prediction interval, simultaneous tolerance intervals may be useful.

Both prediction intervals and simultaneous tolerance intervals, pertain to statements concerning future values of Y's for given x's. The reverse problem of making statements about the x's from which the observed future Y's were obtained is often referred to as the problem of discrimination. The discrimination problem is described as follows: The statistician has n pairs of values $(x_1, Y_1)(x_2, Y_2),\ldots,(x_n, Y_n)$ from which he estimates the regression line $\alpha + \beta x$. He now observes K additional observations $Y_1^*, Y_2^*,\ldots,Y_K^*$ for which the corresponding independent variable values $x_1^*, x_2^*,\ldots,x_K^*$ are unknown. The statistician wishes to estimate these values of x and bracket them by means of simultaneous confidence intervals. This problem was first treated by Mandel [4], and another solution was given in Miller [5].

The analogous situation arises for discrimination problems that motivated the construction of simultaneous tolerance intervals instead of joint prediction intervals. Again, from the sample data, a regression

line $a + bx$ may be constructed, but the number K of discriminations to be made from it may be unknown. Even when K is known, it may be so large that the resultant intervals may be too wide. This problem will be referred to as finding unlimited simultaneous discrimination intervals in regression.

Unlimited simultaneous discrimination intervals $[\underline{D}_{Y*}(P), \bar{D}_{Y*}(P)]$ are sought which are based upon the same estimated linear regression and which have the property that at least 100P percent of the discrimination intervals will contain the true x's with confidence $1-\alpha$. Thus if for a single regression line one asserts that at least 100P percent of the discrimination intervals will contain the correct x's, and similar statements are made repeatedly for different regression lines, then for $100(1-\alpha)$ percent of the different regression lines the statements will be correct. For the other fraction (100α percent) the percentage of discrimination intervals enclosing their true x's may be greater or less than 100P percent for each line.

The use of unlimited simultaneous discrimination intervals is particularly important in bioassay where a standard curve is constructed on which all future assays (discriminations) are to be run. This is similar to the problem of calibrating a measuring instrument where the estimated calibration line is used to correct future readings taken with the instrument. The calibration problem frequently arises in the physical sciences and engineering.

An example of a bioassay to which unlimited simultaneous discrimination techniques could be applied is an assay for immunoglobulins in human sera. Human serum contains three known antibodies (gamma globulins):

γ_A, γ_G, and γ_M. These are routinely assayed in a number of laboratories by the method of radial immunodiffusion.

A radial immunodiffusion assay is performed as follows. An agar plate is prepared with a gel surface containing antiserum to the immunoglobulin to be assayed. Holes or wells are punched into the agar gel. Each plate will have a convenient number of symmetrically placed wells (for example, 6, 12, or 18). A fixed amount of human serum is then pipetted into a well. The antigen in the serum diffuses out from the well and reacts (precipitates) with the antibody in the antiserum gel. A visible precipitate ring is formed about the well. Each well will produce one assay. This is illustrated in Fig. 1 for four assays on a plate with six holes.

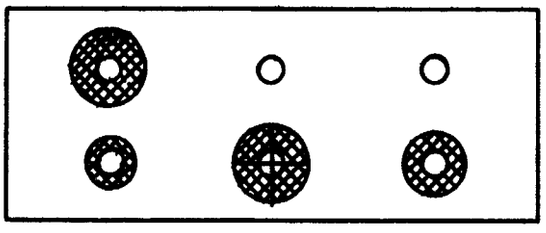


Figure 1

The higher the concentration of the immunoglobulin in the serum the larger the precipitate ring will be. Thus the size of the precipitate ring is a function of the immunoglobulin concentration (and random error).

The size of the ring is customarily measured in one of two ways. The simplest measure is the average of the horizontal and vertical diameters (see Fig. 1). This is the measure we will use, and it is abbreviated by RSD (for ring size diameter). Usually this is quite

satisfactory because only in rare instances will be precipitate rings look noncircular. An alternative method of measuring the ring size is to magnify and silhouette the ring on a piece of paper by means of a light, cut out the silhouetted ring on the paper, and weigh the paper. This method takes much longer and also suffers from inaccuracies.

The standard curve for the bioassay can be constructed from purified immunoglobulin whose concentration has been accurately determined by other means (microkjeldahl analysis and spectrophotometry, for example). Various dilutions of the purified immunoglobulin are assayed on agar plates to find what RSD each dilution gives. From the known concentrations of the dilutions and the resulting RSD's the standard curve can be fitted.

The standard curve of RSD vs. concentration can have various shapes depending on the immunoglobulin, the time allotted for the precipitation, etc. In most cases a reasonable linear fit can be obtained over a wide range of concentrations (or RSD's) in either the original scale or in log scales. In the numerical example to be presented later in Section 4 the assay curve for γ_G is linear (over a large range) for RSD vs. log concentration. Some investigators have found the concentration to be linear vs. the area of the precipitate ring (see [3]). From the statistical point of view the analysis is simplest when a scale for RSD or concentration or both can be found for which the relationship is linear (and normality and homoscedasticity are achieved).

Radial immunodiffusion can be applied to other proteins (for example, serum albumin) besides immunoglobulins. Mancini, Carbonara, and Heremans [3] give a much fuller discussion of the use of this assay than the brief discription here.

Currently at Stanford this assay is being used to study the average levels and variability of the three immunoglobulins in normal, healthy subjects. Once the "normal" levels have been established, this assay and unlimited simultaneous discrimination intervals can be used to pick out "abnormal" subjects. Unlimited simultaneous discrimination intervals can give the clinician an idea of the accuracy and stability of the assayed immunoglobulin levels in patients.

In this paper two techniques for obtaining unlimited simultaneous discrimination intervals will be given. Section 2 presents a procedure which is obtained through the Bonferroni inequality and is briefly described in Miller [5]. Section 3 presents an alternative method based upon an idea presented in the Lieberman-Miller [2] paper and which uses critical points tabled in that paper. Both methods lead to unlimited simultaneous discrimination intervals with the property that at least 100P percent of the discrimination intervals will contain the true x's with confidence at least $1-\alpha$. Section 4 contains a numerical example of the bioassay described earlier. Section 5 presents a discussion and comparison of the two methods for finding unlimited simultaneous discrimination intervals.

Throughout this paper it will be assumed that

$$(1) \qquad Y_i = \alpha + \beta x_i + e_i, \qquad i = 1,\ldots,n ,$$

where the e_i are independent $N(0,\sigma^2)$. The pairs (x_i,Y_i) are the original observations on the regression line. We will use the following customary estimators of α, β, σ^2:

$$\hat{\alpha} = a = \bar{Y} - b\bar{x} ,$$

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} ,$$

(2)

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - a - bx_i)^2 .$$

The distribution theory of a, b, s^2 is so well-known that there is no need to summarize it here.

Future observations on the dependent variable will be designated by a superscript *, that is, Y^*. Let μ_{Y^*} denote the true mean value associated with Y^*, and x^* the value of the independent variable connected with μ_{Y^*} (i.e., $\mu_{Y^*} = \alpha + \beta x^*$). For a future observation Y^* the customary estimator of the value x^* is $\hat{x}^* = (Y^* - a)/b$. This paper is concerned with constructing upper and lower limits for each x^* no matter how many future Y^* are observed.

II. Bonferroni Intervals

The simple Bonferroni inequality says

(3) $$P(A \cap B) \geq 1 - P(A^c) - P(B^c) .$$

This inequality is useful in combining two confidence statements for if A is one statement with confidence $1-(\alpha/2)$ and B is another with confidence $1-(\alpha/2)$, then both statements hold with confidence at least $1-(\alpha/2)-(\alpha/2) = 1-\alpha$.

The idea behind the Bonferroni intervals is to combine (for a given Y^*) the confidence interval on μ_{Y^*} with the confidence band on the line $\mu = \alpha + \beta x$. If the standard deviation σ were known, then each interval

(4) $$Y^* \pm N(P)\sigma \ ,$$

where $N(P)$ is defined by

(5) $$P = \frac{1}{\sqrt{2\pi}} \int_{-N(P)}^{+N(P)} e^{-y^2/2} dy \ ,$$

would have probability P of containing the true μ_{Y^*}. For each μ in the interval (4), the Working-Hotelling confidence band on the line $\mu = \alpha + \beta x$ gives a confidence interval for the corresponding $x = (\mu-\alpha)/\beta$. The union of all these confidence intervals as μ varies over the interval (4) would give a discrimination interval for x^*.

But σ is not known. However, if it can be bounded above with known confidence, then this bound can be inserted in (4) to produce intervals which have probability at least P of containing their true μ_{Y^*}'s. The Bonferroni inequality is used to combine the confidence in the bound on σ with the confidence on the band about the line $\mu=\alpha+\beta x$.

This procedure is illustrated in Fig. 2. The interval in the brace on the y-axis is the confidence interval for μ_{Y^*}; the interval in the brace on the x-axis is the discrimination interval. If μ_{Y^*} is contained in the interval $Y^* \pm \Delta$, and <u>if</u> the Working-Hotelling confidence band contains the line $\mu = \alpha + \beta x$, <u>then</u> the point $(x^*, \alpha + \beta x^*)$ must lie in the shaded region and x^* must lie in the indicated (discrimination) interval.

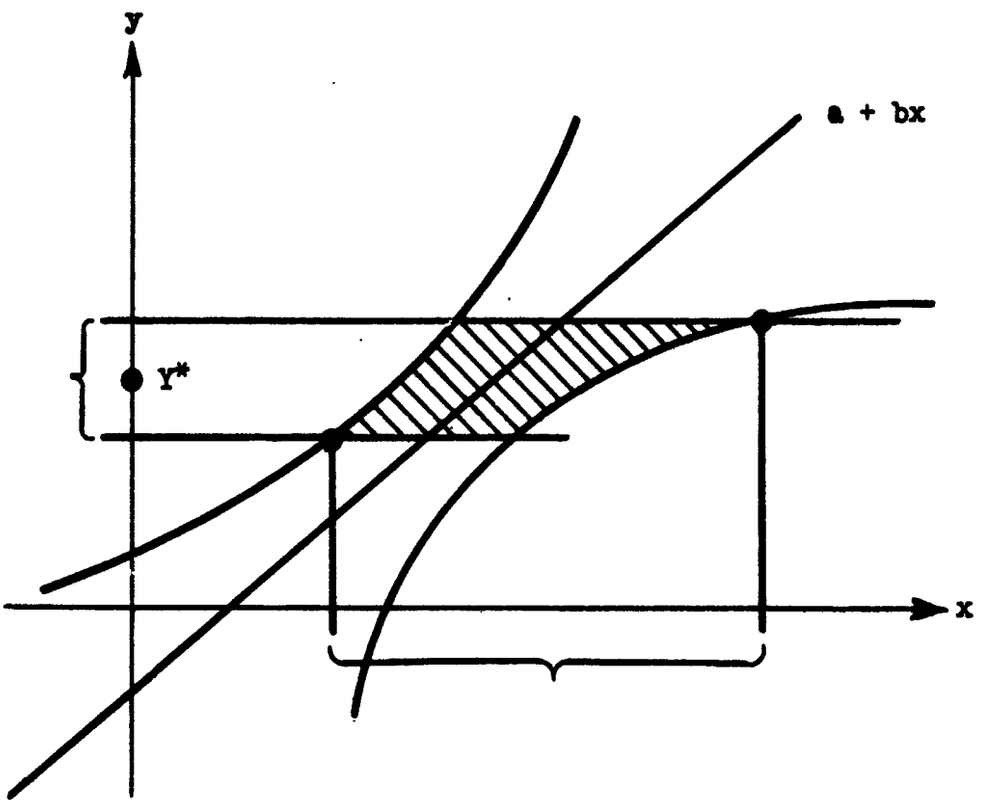


Figure 2

Mathematically the discrimination intervals are constructed as follows. The Working-Hotelling confidence band on the regression line $\mu = \alpha + \beta x$ is

$$(6) \qquad \alpha + \beta x \in a + bx \pm \left(2F^{\alpha/2}_{2,n-2}\right)^{\frac{1}{2}} s\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_1 (x_i-\bar{x})^2}\right)^{\frac{1}{2}}$$

for all x. The critical point $F^{\alpha/2}_{2,n-2}$ is the upper $\alpha/2$ percentile point of the F distribution with degrees of freedom $2,n-2$. With probability $1-(\alpha/2)$ the band (6) contains the true regression line

The unknown standard deviation σ can be bounded above by

$$(7) \qquad \sigma \leq \left(\frac{n-2}{\alpha/2\ \chi^2_{n-2}}\right)^{\frac{1}{2}} s$$

where $_{\alpha/2}\chi^2_{n-2}$ is the lower $\alpha/2$ percentile point of the χ^2 distribution with $n-2$ degrees of freedom. With probability $1-(\alpha/2)$ the inequality (6) holds true.

By the Bonferroni inequality (3), the probability (with respect to $(x_1,Y_1),\ldots,(x_n,Y_n)$) that both (6) and (7) hold is at least $1-\alpha$.

For a future observation Y^* the 100P percent confidence interval on its true mean μ_{Y^*} is $Y^* \pm N(P)\sigma$. As the number of future observations tends to infinity the proportion of intervals which correctly contain their true means converges to P by the law of large numbers. But the interval $Y^* \pm N(P)\sigma$ is contained in the interval

$$(8) \qquad Y^* \pm N(P)\left(\frac{n-2}{{}_{\alpha/2}\chi^2_{n-2}}\right)^{\frac{1}{2}} s$$

(with probability $1-(\alpha/2)$). Thus, the (limiting) proportion of intervals (8) which correctly contain their true means is at least P (with probability $1-(\alpha/2)$).

The discrimination interval is obtained by intersecting the interval (8) on the y-axis with the confidence band on $\alpha + \beta x$, and projecting the intersection onto the x-axis. This is illustrated in Fig. 2. Actually, Fig. 2 represents the nice case. The discrimination "interval" can be the entire real line or the union of two semi-infinite intervals. These pathological cases will occur when the regression line is too flat (i.e., when b is too near zero relative to its variability). These cases are directly analogous to what can happen for a single discrimination where this pathology is referred to as the Fieller-Creasy paradox. The reader can visualize when these pathological cases occur by redrawing

Fig. 2 with a much flatter slope and shifting the μ_{Y^*} interval up and down. The necessary and sufficient condition for the discrimination "interval" to be a nice finite interval is

$$b^2 > \frac{2F^{\alpha/2}_{2,n-2}\, s^2}{\sum_i (x_i-\bar{x})^2} \quad . \tag{9}$$

Only the nice case will be considered in this paper.

Let $[\underline{D}_{Y^*}(P),\ \bar{D}_{Y^*}(P)] = [\underline{B}_{Y^*}(P),\ \bar{B}_{Y^*}(P)]$ be the discrimination interval (B for Bonferroni). If the sample regression line has positive slope, the upper endpoint $\bar{B}_{Y^*}(P)$ is the root of the equation

$$\begin{aligned} a + bx^* - \left(2F^{\alpha/2}_{2,n-2}\right)^{\frac{1}{2}} s \left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)^{\frac{1}{2}} \\ = Y^* + N(P)\left(\frac{n-2}{\alpha/2\ \chi^2_{n-2}}\right)^{\frac{1}{2}} s \ , \end{aligned} \tag{10}$$

and $\underline{B}_{Y^*}(P)$ is the root of the equation

$$\begin{aligned} a + bx^* + \left(2F^{\alpha/2}_{2,n-2}\right)^{\frac{1}{2}} s \left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)^{\frac{1}{2}} \\ = Y^* - N(P)\left(\frac{n-2}{\alpha/2\ \chi^2_{n-2}}\right)^{\frac{1}{2}} s \ . \end{aligned} \tag{11}$$

These roots are:

$$\bar{B}_{Y^*}(P) = \bar{x} + \frac{b(Y^*-\bar{Y}+Q) + R^{\frac{1}{2}}[(Y^*-\bar{Y}+Q)^2 + (\sum_i (x_i-\bar{x})^2/n)(b^2-R)]^{\frac{1}{2}}}{b^2-R} \tag{12}$$

and

$$\underline{B}_{Y^*}(P) = \bar{x} - \frac{b(Q-Y^*+\bar{Y}) - R^{\frac{1}{2}}[(Q-Y^*+\bar{Y})^2 + (\sum_i (x_i-\bar{x})^2/n)(b^2-R)]^{\frac{1}{2}}}{b^2-R} \tag{13}$$

where

$$R = (2F^{\alpha/2}_{2,n-2}\, s^2)/\sum_i (x_i - \bar{x})^2$$

$$Q = N(P)[(n-2)/{}^{\alpha/2}\chi^2_{n-2}]^{\frac{1}{2}}\, s\ .$$

If the sample regression line has negative slope, then $\bar{B}_{Y*}(P)$ is the root of the equation (11) and $\underline{B}_{Y*}(P)$ is the root of the equation (10).

III. Augmented F Intervals

The intervals in the preceding section were derived from the joint probability statement

$$P\{\alpha + \beta x \in a + bx \pm (2F^{\alpha/2}_{2,n-2})^{\frac{1}{2}}\, s \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)^{\frac{1}{2}} \text{ for all } x, \quad \underline{\text{and}} \quad \sigma \le \left(\frac{n-2}{\alpha/2\chi^2_{n-2}}\right)^{\frac{1}{2}} s\} \ge 1-\alpha\ , \tag{14}$$

whose validity stemmed from the Bonferroni inequality. In this section a similar joint probability statement will be obtained from a different approach.

In [2] Lieberman and Miller proved that

$$P\{|(a-\alpha) + (b-\beta)x \pm N(P)\sigma| \le c^* s \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2} + N^2(P)\right)^{\frac{1}{2}} \text{ for all } x, P\} = 1-\alpha\ , \tag{15}$$

where the critical point c^* is defined by

$$(16) \qquad P\left\{\frac{Z_1^2 + Z_2^2 + 1}{\chi^2_{n-2}/(n-2)} \leq (c^*)^2\right\} = 1 - \alpha .$$

The random variables Z_1, Z_2 are independent $N(0,1)$, and χ^2_{n-2} is a χ^2 variable with $n-2$ degrees of freedom which is independent of Z_1 and Z_2. For want of a name the statistic

$$(17) \qquad \frac{Z_1^2 + Z_2^2 + 1}{\chi^2_{n-2}/(n-2)}$$

will be referred to as the augmented F statistic. Tables of c^* for $\alpha = .5, .3, .1, .05, .01, .001$; $n-2 = 1(1)\ 30(5)\ 50(10)\ 100$ are given in [2].

By taking $N(P) = 0$ for the first inequality, and letting $N(P) \to +\infty$ for the second inequality, the expression (15) implies that

$$(18) \qquad P\{|(a-\alpha) + (b-\beta)x| \leq c^* s \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)^{\frac{1}{2}} \text{ for all } x ,$$
$$\underline{\text{and}} \quad \sigma \leq c^* s\} \geq 1 - \alpha .$$

The probability in (18) is actually greater than $1-\alpha$ because not all combinations of x and $N(P)$ are utilized inside the probability sign. Thus, the confidence band

$$(19) \qquad \alpha + \beta x \in a + bx \pm c^* s \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)^{\frac{1}{2}} \text{ for all } x ,$$

and the bound

$$(20) \qquad \sigma \leq c^* s ,$$

hold simultaneously with probability at least $1-\alpha$.

The construction of the discrimination intervals now proceeds exactly as in the preceding section with (19) replacing (6) and (20) replacing (7). The discrimination "intervals" will all be nice finite intervals if and only if

$$b^2 > \frac{(c^*)^2 s^2}{\sum_i (x_i - \bar{x})^2} . \tag{21}$$

Only the nice case will be considered in this paper.

Let $[\underline{D}_{Y^*}(P), \overline{D}_{Y^*}(P)] = [\underline{C}_{Y^*}(P), \overline{C}_{Y^*}(P)]$ be the discrimination interval (C for the method based upon c^*). If the sample regression line has positive slope, the upper endpoint $\overline{C}_{Y^*}(P)$ is the root of the equation

$$\begin{aligned} a + bx^* - c^* s \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)^{\frac{1}{2}} \\ = Y^* + N(P)\, c^*\, s , \end{aligned} \tag{22}$$

and $\underline{C}_{Y^*}(P)$ is the root of the equation

$$\begin{aligned} a + bx^* + c^* s \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)^{\frac{1}{2}} \\ = Y^* - N(P)\, c^*\, s . \end{aligned} \tag{23}$$

These roots are

$$\overline{C}_{Y^*}(P) = \bar{x} + \frac{b(D + Y^* - \bar{Y}) + E^{\frac{1}{2}}[(D + Y^* - \bar{Y})^2 + (\sum_i (x_i - \bar{x})^2/n)(b^2 - E)]^{\frac{1}{2}}}{b^2 - E} \tag{24}$$

and

$$\underline{C}_{Y^*}(P) = \bar{x} - \frac{b(D - Y^* + \bar{Y}) - E^{\frac{1}{2}}[(D - Y^* + \bar{Y})^2 + (\sum_i (x_i - \bar{x})^2/n)(b^2 - E)]^{\frac{1}{2}}}{b^2 - E} \tag{25}$$

where

$$E = (c^{*2}\ s^2)/\sum_i (x_i - \bar{x})^2$$

$$D = N(P)\ c^{*2}\ s\ .$$

If the sample regression line has negative slope, then $\bar{C}_{Y*}(P)$ is the root of the equation (23) and $\underline{C}_{Y*}(P)$ is the root of the equation (22).

IV. Numerical Example

A biassay of the gamma globulin γ_G was performed at the Stanford University Medical Center using the method of radial immunodiffusion described in section I. The ring size diameter (RSD) was measured on two wells for each of seven known concentrations of γ_G. The resulting data are given in Table 1.

Table 1

y(R.S.D.)[1]	z(mg % γ_G)	x($\log_{10}$z)
68 ,68	1383.6	3.1410
62 ,62	696.8	2.8431
62.5,62.5	716.9	2.8555
55.5,55.5	328.3	2.5163
56 ,56	335.0	2.5250
48 ,49	147.4	2.1685
48 ,48	140.7	2.1483

$\bar{x} = 2.5997$; $\bar{y} = 57.2143$; $a = 4.8790$; $b = 20.1300$;

$s = .2569$; $\sum_{i=1}^{14} (x_i - \bar{x})^2 = 1.6398$

[1] Each of the fourteen measurements of the y's is the average of the horizontal and vertical diameters.

The usual test of linearity shows that a simple linear regression of RSD on log concentration fits the data quite well. The graph of Fig. 3 presents the above data and the associated estimated regression line $\hat{y} = a + bx$.

At each of three hypothetical future values of Y^*, 57.20, 70, and 80, unlimited simultaneous discrimination intervals for the associated x^* values have been computed at $\alpha = .01, .05$ and $P = .30, .80$ using the Bonferroni and augmented F methods (equations (12), (13) and (24), (25)). The results of these computations are presented in Table 2.

Table 2

α	P	Y^*(RSD)	$\underline{B}_{Y^*}(P), \overline{B}_{Y^*}(P)$	$\overline{B}_{Y^*}(P) - \underline{B}_{Y^*}(P)$	$\underline{C}_{Y^*}(P), \overline{C}_{Y^*}(P)$	$\overline{C}_{Y^*}(P) - \underline{C}_{Y^*}(P)$
.01	.30	57.20	2.582,2.617	.035	2.566,2.633	.067
		70	3.203,3.269	.066	3.188,3.285	.097
		80	3.682,3.786	.104	3.667,3.801	.134
	.80	57.20	2.575,2.624	.049	2.520,2.679	.159
		70	3.196,3.277	.081	3.144,3.333	.189
		80	3.675,3.794	.119	3.623,3.849	.226
.05	.30	57.20	2.585,2.614	.029	2.574,2.625	.051
		70	3.209,3.262	.053	3.199,3.273	.074
		80	3.692,3.774	.082	3.682,3.785	.103
	.80	57.20	2.577,2.622	.045	2.538,2.660	.122
		70	3.210,3.271	.069	3.164,3.310	.146
		80	3.684,3.783	.099	3.648,3.821	.173

The Y^* value 57.20 is equal to $a+bx$. Therefore, for fixed α and P, the intervals for the corresponding x^* have lengths shorter than intervals corresponding to any other value of Y^*.

Because of the symmetry of this problem about the point $Y^* = a+b\bar{x}$, an interval for x^* corresponding to $y^* = 44.40$ would have the same

length as the interval for x^* corresponding to $y^* = 70$. The same relationship exists for intervals on x^*'s corresponding to Y^* values of 34.40 and 80. It is implicit here that a Y^* value as large as 80 (or as small as 34.40) is within the range of linearity.

This example indicates that both methods can be of practical importance. The lengths of the intervals are short enough to be useful for picking out subjects who have "abnormal" concentrations of γ_G (see section I). It is also evident that for this example the Bonferonni intervals are shorter than the intervals obtained using the augmented F procedure.

V. Comparisons and Discussion

The equations for the Bonferroni intervals (Sec.2) and the augmented F intervals (Section 3) are not related in a way that enables one to completely define conditions under which one method always produces shorter intervals than the others. Of course, the interval lengths can always be compared for given data. In addition, the interval lengths for the two methods are amenable to comparison when the Y^* values are close to the number $a+b\bar{x}$ or when the Y^* values are far from $a+b\bar{x}$. These two situations are discussed below.

The comparison for the former situation is stated in terms of the statistic $\frac{b^2\sum(x_i-\bar{x})^2}{s^2} = f$, say. Distribution theory for f is well-known; in fact, f has an F distribution with 1 and $n-2$ degrees of freedom under the hypothesis that $\beta = 0$ and a non-central F-distribution when $\beta \neq 0$. Hence, it can be said that if f is quite large, then b is highly significantly different from zero, and conversely.

Notice that f is greater than $2F^{\alpha/2}_{2,n-2}$ and f is greater than c^{*2} if and only if all Bonferroni and augmented F intervals are of finite length (see equations 9 and 21). To avoid infinite-length intervals, let $f > 2F^{\alpha/2}_{2,n-2}$ and $f > c^{*2}$ for the discussion below.

Suppose Y^* is so close to $a+b\bar{x}$ that $Y^*-(a+b\bar{x})$ makes a negligible contribution to the computations of $B_{Y^*}(P)$ and $C_{Y^*}(P)$. Then,

(26)

$$\bar{C}_{Y^*}(P) - \underline{C}_{Y^*}(P) < \bar{B}_{Y^*}(P) - \underline{B}_{Y^*}(P) \qquad \text{if and only if}$$

$$h(f) \equiv f - 2F^{\alpha/2}_{2,n-2}$$

$$< (f-c^{*2})\left\{\frac{N(P)\left(\frac{n(n-2)f}{\chi^2_{\alpha/2,n-2}}\right)^{\frac{1}{2}} + \left[nN^2(P)\frac{(n-2)2F^{\alpha/2}_{2,n-2}}{\chi^2_{\alpha/2,n-2}} - (2F^{\alpha/2}_{2,n-2})^2 + 2fF^{\alpha/2}_{2,n-2}\right]^{\frac{1}{2}}}{N(P)\ c^*(nf)^{\frac{1}{2}} + [nN^2(P)\ c^{*4} - c^{*4} + c^{*2}f]^{\frac{1}{2}}}\right\} \equiv g(f).$$

By comparing tabled values of c^* and $F^{\alpha/2}_{2,n-2}$, one finds that $c^{*2} < 2F^{\alpha/2}_{2,n-2}$ at $\alpha = .001, .01, .05, .10,$ and $.30$, and $n = 3, 7, 12, 32, 62,$ and 102. It is reasonable to state without formal proof that at least for $3 \le n \le 102$ and $.001 \le \alpha \le .30$, $c^{*2} < 2F^{\alpha/2}_{2,n-2}$. Since $g(f) \ge 0$ for $f \ge c^{*2}$ and $h(f) = 0$ at $f = 2F^{\alpha/2}_{2,n-2}$, then $h(f) \le g(f)$ and $\bar{C}_{Y^*}(P) - \underline{C}_{Y^*}(P) \le \bar{B}_{Y^*}(P) - \underline{B}_{Y^*}(P)$, when $2F^{\alpha/2}_{2,n-2} \le f \le f_0$, where f_0 is the point at which the curves $h(f)$ and $g(f)$ first cross above $2F^{\alpha/2}_{2,n-2}$; more explicitly, where f_0 is the infimum of f values greater than $2F^{\alpha/2}_{2,n-2}$ such that $g(f) < h(f)$.

Some representative values of f_0 for various levels of α, P, and n are given in Table 3. These values were found using an iterative procedure.

After computing f_0 for the desired α and P levels, one can find $P(2F^{\alpha/2}_{2,n-2} < f < f_0)$ for several values of β. This information

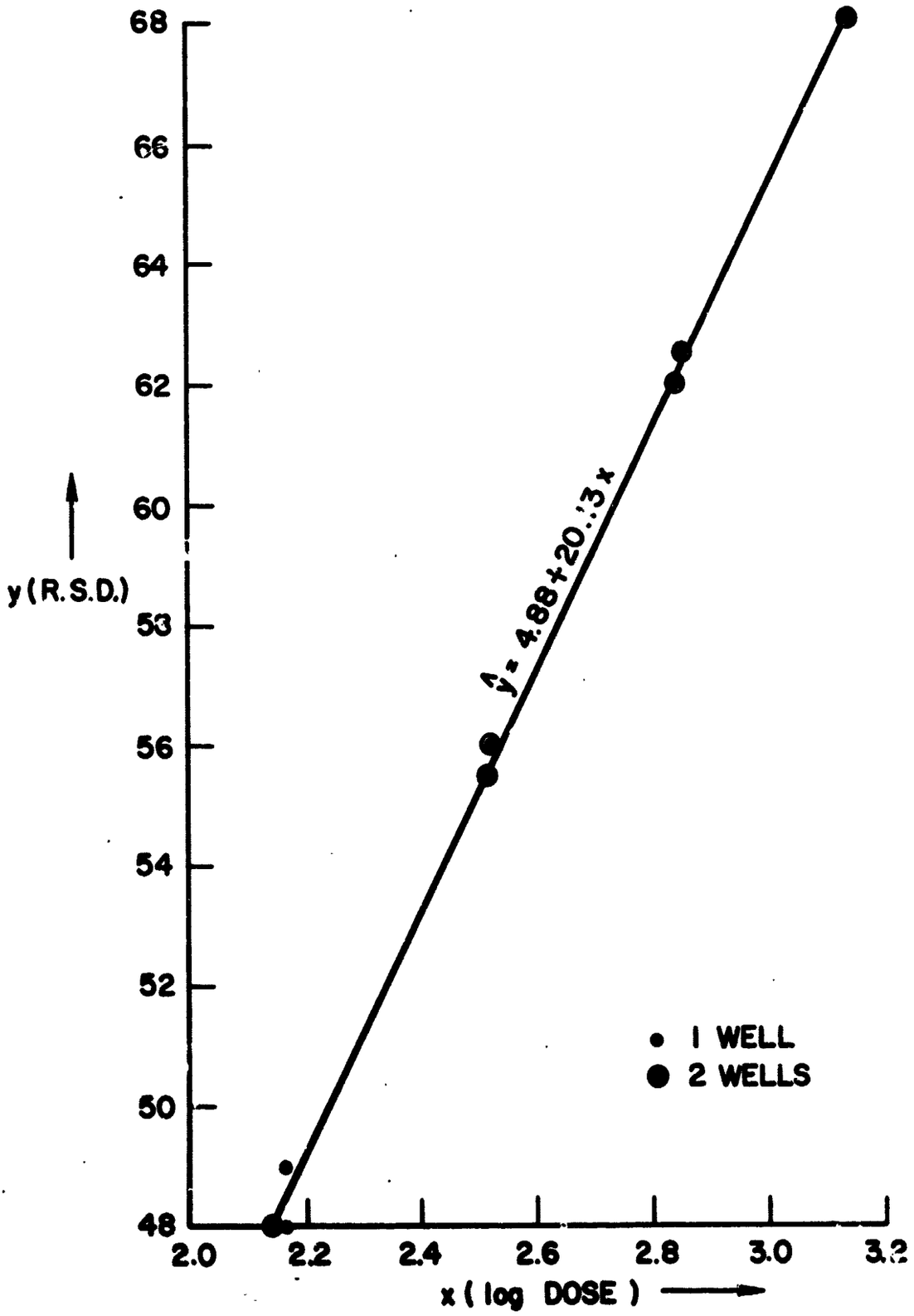


Figure 3. A simple linear regression estimate based on measurements of ring size diameter on two wells at each of 7 concentrations of gamma globulin (mg. % of γ_g) using a radial immunodiffusion assay.

can be used by the statistician who must recommend an unlimited simultaneous discrimination interval method to his client before the data are observed. For example, if the probabilities are large, the statistician can recommend the use of the augmented F method knowing that it will produce shorter intervals than the Bonferroni method for Y^* values near $a+b\bar{x}$ most of the time.

Table 3

n	α	P	$2F^{\alpha/2}_{2,n-2}$	c^{*2}	f_0
7	.10	.50,.80,.90	11.6	10.3	14.2
7	.30	.50,.80,.90	5.7	4.8	8.4
15	.001	.50,.80,.90	29.0	27.6	30.1
15	.10	.50,.80,.90	7.6	7.0	8.4
15	.30	.50,.80,.90	4.4	3.9	5.4
102	.001	.50,.80,.90	16.4	16.0	16.6
102	.10	.50,.80,.90	6.2	5.7	6.5

Although $2F^{\alpha/2}_{2,n-2} < f < f_0$ implies that $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P) < \bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$, it is not necessarily true that $f > f_0$ implies that $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P) > \bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$. This is because the curves $h(f)$ and $g(f)$ may cross more than once. It can be shown that if f is very large and $N(P)\sqrt{n}$ is large enough then $h(f) > g(f)$ and $\bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P) < \bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P)$. A "large enough" value for $N(P)\sqrt{n}$ is not a very restrictive condition; e.g., $N(P)$ must be greater than .08 if $\alpha = .10$, $n = 102$, $N(P)$ must be greater than 1.06 if $\alpha = .001$, $n = 12$. Thus it is generally true if b is highly significantly different from zero and if Y^* is not far from $a+b\bar{x}$, then the Bonferroni method will produce shorter intervals than the augmented F method.

For the example of section 4, $f = 10,068$ indicating that b is highly significantly different from zero. It is not surprising, therefore, that every Bonferroni interval in Table 1 is shorter than the corresponding augmented F interval.

Now consider the situation where the Y^* values are such that $|Y^* - (a+b\bar{x})|$ is so large that $\frac{s^2}{n}(f-2F^{\alpha/2}_{2,n-2})$, $\frac{s^2}{n}(f-c^{*2})$, $N(P)[\frac{n-2}{\alpha/2\chi^2_{n-2}}]^{\frac{1}{2}}s$, $N(P)$ c^*s are negligible in comparison. Then $\bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$ is approximately equal to

$$2b\left(\frac{2F^{\alpha/2}_{2,n-2}}{f}\right)^{\frac{1}{2}} |Y^* - (a+b\bar{x})|$$

and $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P)$ is approximately equal to

$$2b\,\frac{c^*}{\sqrt{f}}\,|Y^* - (a+b\bar{x})| \ .$$

Therefore, $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P) < \bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$ if and only if $c^{*2} < 2F^{\alpha/2}_{2,n-2}$. As indicated earlier, $2F^{\alpha/2}_{2,n-2} > c^{*2}$ for $3 \le n \le 102$ and $.001 < \alpha < .30$. Consequently, for Y^* values far from $a+b\bar{x}$, it is generally true that the augmented F method provides shorter intervals than does the Bonferroni method.

For the immunodiffusion data discussed in Section 4,

$$\bar{B}_{Y^*}(.80)-\underline{B}_{Y^*}(.80) = .023486 + .0015828\,\{[(Y^*-(a+b\bar{x}) + .23615)^2 + 47.415]^{\frac{1}{2}} + [(.23615-Y^* + (a+b\bar{x}))^2 + 47.415]^{\frac{1}{2}}\}$$

and

$$\bar{C}_{Y^*}(.80)-\underline{C}_{Y^*}(.80) = .10088 + .0015263\,\{[(1.01440 + Y^* - (a+b\bar{x}))^2 + 47.418]^{\frac{1}{2}} + [(1.01440 - Y^* + (a+b\bar{x}))^2 + 47.418]^{\frac{1}{2}}\},$$

if $P = .80$, $\alpha = .05$. Hence, $\bar{C}_{Y^*}(.80)-\underline{C}_{Y^*}(.80) < \bar{B}_{Y^*}(.80)-\underline{B}_{Y^*}(.80)$

if and only if $|Y^*-(a+b\bar{x})| > 686$. Since Y^* cannot be negative, the augmented F intervals will be shorter than the Bonferroni intervals if and only if the Y^* values are greater than 744.

Obviously, an RSD value of 744 or larger will never be observed and one concludes that, if $P = .80$ and $\alpha = .05$, the Bonferroni method is definitely better than the augmented F method for estimating unlimited simultaneous discrimination intervals based on the immunodiffusion data.

Extremely large values of Y^* are not always necessary for one to be certain that $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P)$ is less than $\bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$. If f is small, n is small, α is large, and P is small then $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P)$ may be less than $\bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$ for all Y^* values greater than a point which is not far from $a+b\bar{x}$. For any of the usual choices of α and P levels and normal regression data (n and f not small), however, it appears that Y^* would have to be very far from $a+b\bar{x}$ before one could be certain that $\bar{C}_{Y^*}(P)-\underline{C}_{Y^*}(P) < \bar{B}_{Y^*}(P)-\underline{B}_{Y^*}(P)$.

It should be mentioned that the augmented F method can be expected to produce short intervals when n is small. In fact, in the limiting case of $n = 3$, the augmented F procedure yields shorter intervals than the Bonferroni method for any choices of α, P, Y^*, and any data (subject to $f > c^{*2}$). This is because at $n = 3$, $c^{*2} < \frac{n-2}{\alpha/2\chi^2_{n-2}}$ and $c^{*2} < 2F^{\alpha/2}_{2,n-2}$ for every α.

The above comparisons of the lengths of intervals provided by the Bonferroni and augmented F procedures do not yield explicit results. Nevertheless, one has the impression that in most problems where these methods can be useful, the method based on the Bonferroni inequality will yield shorter unlimited simultaneous discrimination intervals--especially when the future Y^* values are not expected to be far from $a+b\bar{x}$.

Acknowledgment

The authors would like to thank Judith Grindle for programming the computations presented in this paper.

Bibliography

[1] Lieberman, G. J., 1961. Prediction Regions for Several Predictions from a Single Regression Line. Technometrics, 3, 21-27.

[2] Lieberman, G. J., and Miller, R. G., 1963. Simultaneous Tolerance Intervals in Regression. Biometrika, 50, 155-168.

[3] Mancini, G., Carbonara, A. O., and Heremans, J. F., 1965. Immunochemical Quantitation of Antigens by Single Radial Immunodiffusion. Immunochemistry, 2, 235-254.

[4] Mandel, J., 1958. A Note on Confidence Intervals in Regression Problems. Ann. Math. Statist., 29, 903-907.

[5] Miller, R. G., 1966. Simultaneous Statistical Inference. McGraw-Hill Book Co., New York.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Department of Statistics
Stanford University

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b GROUP

3. REPORT TITLE

Unlimited Simultaneous Discrimination Intervals in Regression

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report, July 1966

5. AUTHOR(S) (Last name, first name, initial)

G. J. Lieberman, R. G. Miller, Jr., and M. A. Hamilton

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
July 29, 1966	24	5

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Nonr 225(53)	Technical Report. No. 90
b. PROJECT NO. NR 042-002	
c. Nonr-225(53) (FBM)	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	None

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Logistics and Math. Statistics Branch Office of Naval Research Washington, D.C. 20360

13. ABSTRACT The discrimination problem can be described as follows: The statistician has n pairs of values (x_1,Y_1), (x_2,Y_2), ... , (x_n,Y_n) from which he estimates the regression line $\alpha + \beta x$. He now observes K additional observations $Y_1^*, Y_2^*,\ldots,Y_K^*$ for which the corresponding independent variable values $x_1^*, x_2^*,\ldots,x_K^*$ are unknown. The statistician wishes to estimate these values of x and bracket them by means of simultaneous confidence intervals. This problem was first treated by Mandel (1958) and another solution was given in Miller's book (1966). When K is unknown and possibly arbitrarily large, these results do not apply. A solution to this problem of arbitrary K is given in terms of unlimited simultaneous discrimination intervals. Unlimited simultaneous discrimination intervals $[\underline{D}_{Y^*}(P), \overline{D}_{Y^*}(P)]$ are presented which are based upon the same estimated linear regression and which have the property that at least 100P per cent of the discrimination intervals will contain the true x's with confidence $1 - \alpha$. Thus if for a single regression line one asserts that at least 100P per cent of the discrimination intervals will contain the correct x's, and similar statements are made repeatedly for different regression lines, then for $100(1 - \alpha)$ per cent of the different regression lines the statements will be correct. For the other fraction (100α per cent) the percentage of discrimination intervals enclosing their true x's may be greater or less than 100P per cent for each line. In this paper two techniques for obtaining unlimited simultaneous discrimination intervals are given. The first method is a procedure which is obtained through the Bonferroni inequality, while the second technique is based upon an idea presented in a paper by Lieberman and Miller (1963). A numerical example is analyzed. A general discussion and comparison of the two methods for finding unlimited simultaneous discrimination intervals is given.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Bonferroni inequality confidence intervals discrimination discrimination interval regression unlimited simultaneous discrimination intervals						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.